Zastosowanie uogólnionych modeli liniowych i uogólnionych mieszanych modeli liniowych do analizy danych dotyczących występowania zębiniaków

Wojciech Niemiro, Jacek Tomczyk i Marta Zalewska

Uniwersytet Warszawski i UMK Toruń, Uniwersytet Kardynała Stefana Wyszyńskiego, Warszawski Uniwersytet Medyczny

Konferencja Zastosowań Matematyki
Zakopane-Kościelisko, wrzesień 2016

Plan

1. Dane i modele regresji
   - Dane
   - Cel analizy
   - Modele regresji logistycznej
2. Dalsze modele
   - Modele kwadratowo-logistyczne
   - Regresja wieloraka (wiele zmiennych objaśniających)

Plan

1. Dane i modele regresji
   - Dane
   - Cel analizy
   - Modele regresji logistycznej
2. Dalsze modele
   - Modele kwadratowo-logistyczne
   - Regresja wieloraka (wiele zmiennych objaśniających)

Dane wykopaliskowe

780 zębów, 120 osobników, 6 cech (zmiennych).

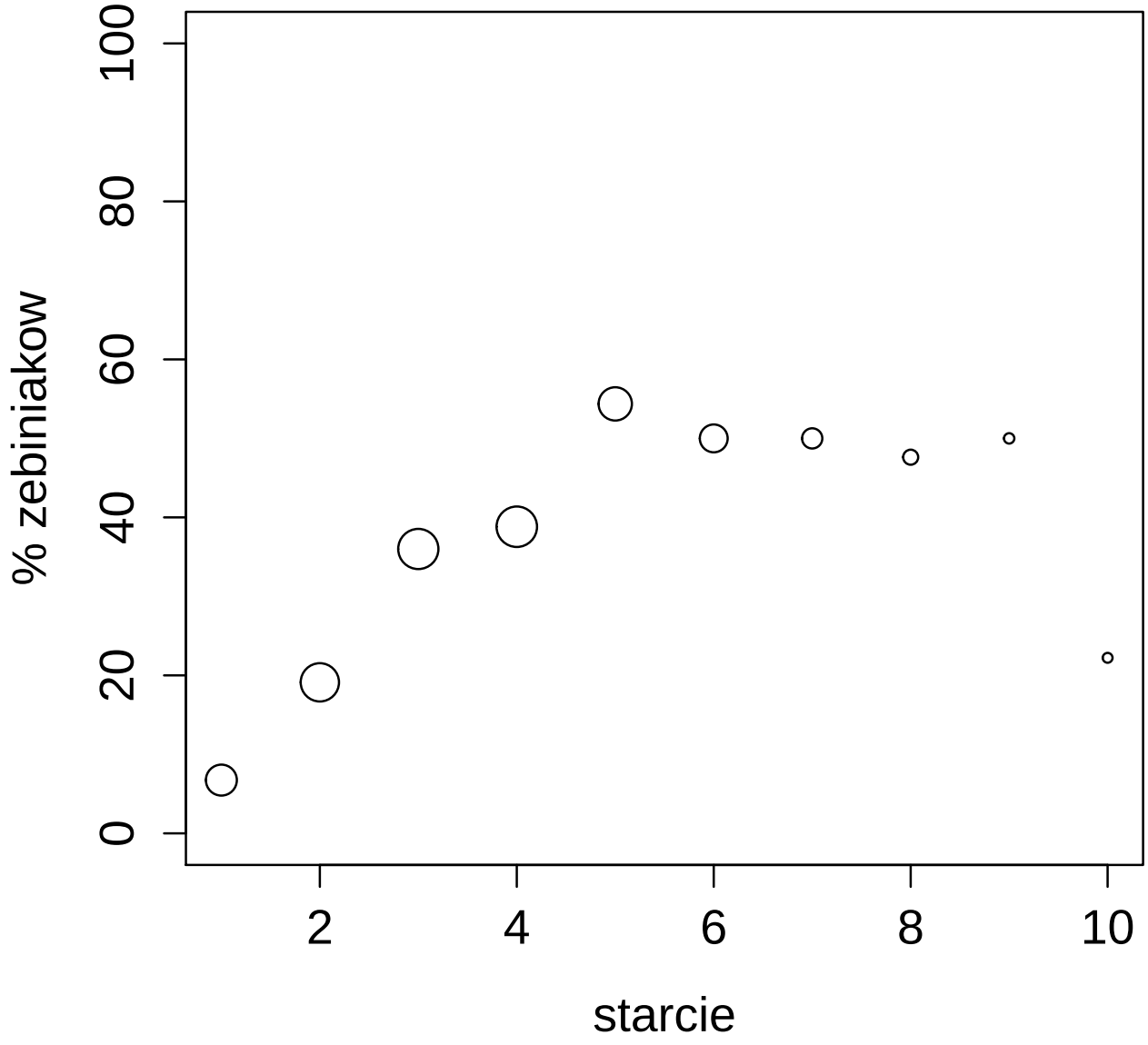
Cecha wyjaśniana: obecność zębiniaków (zmienna „zebiniak” o wartościach 0-1).

	Osobnik	plec	wiek	typ	starcie	prochnica	zebiniak
1	12	0	22	16	3	1	1
2	12	0	22	17	2	1	1
3	12	0	22	18	1	1	1
4	12	0	22	26	4	1	0
5	12	0	22	27	3	1	0
6	12	0	22	28	2	1	0
...	...	...	...	...	...	...	...
780	45	1	50	37	7	0	1

Cel analizy statystycznej

Wyjaśnić/wykryć zależność występowania zębiniaków od innych zmiennych (cech), w szczególności od stopnia starcia zęba.

Zależność zębiniaków od stopnia starcia



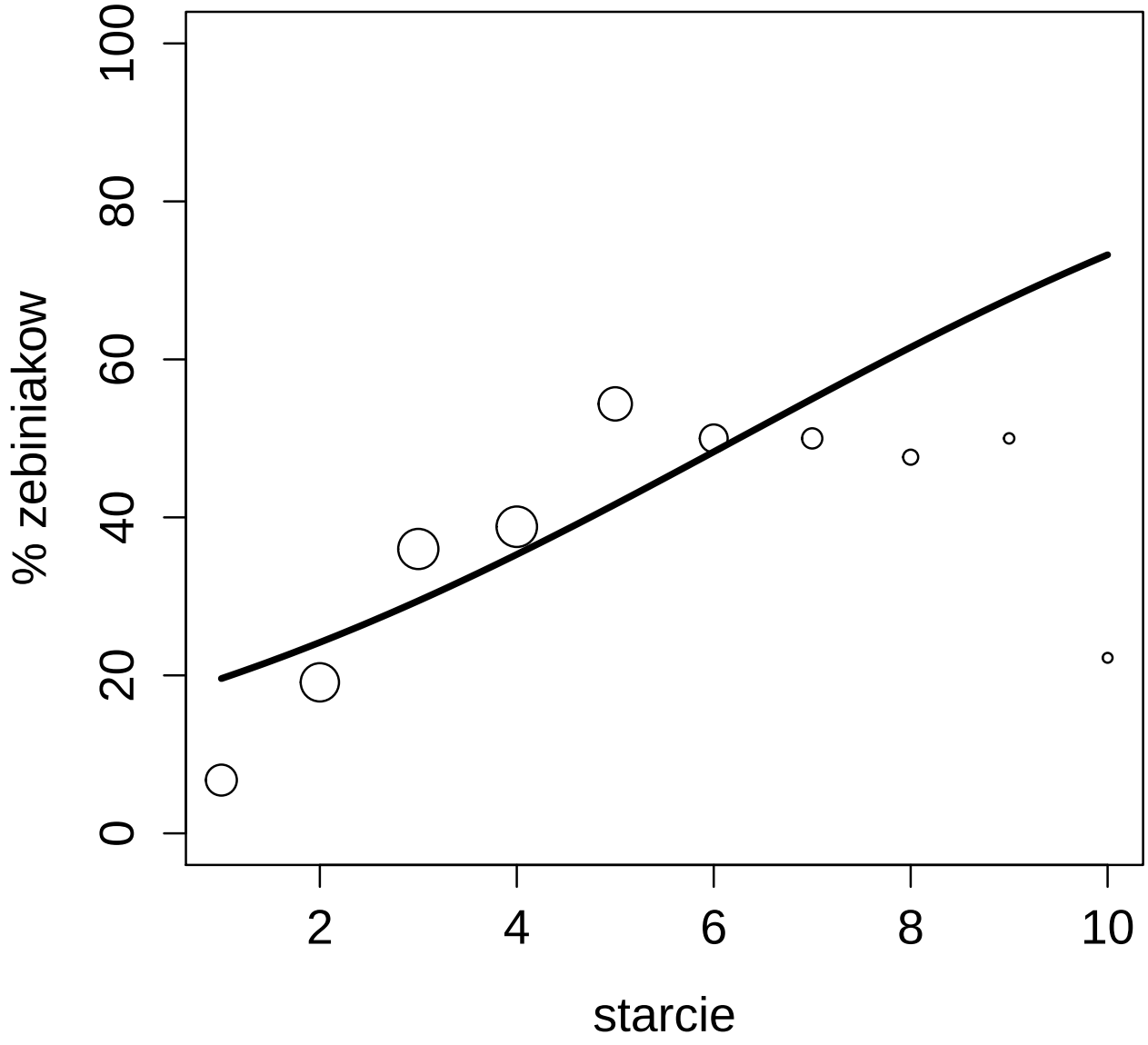
Model regresji logistycznej

$$\log \frac{p_i}{1-p_i} = \beta_0 + x_i^T \beta,$$

$$p_i = \frac{\exp\{\beta_0 + x_i^T \beta\}}{1 + \exp\{\beta_0 + x_i^T \beta\}},$$

- p_i – prawdopodobieństwo wystąpienia zębiniaka.
- x_i – wektor zmiennych „wyjaśniających”: podzbiór cech (plec, wiek, typ, starcie, prochnica).
- $\beta_0 + x_i^T \beta$ – liniowy predyktor.

Regresja logistyczna



Wątpliwości i zarzuty

- Czy model jest adekwatny?
- Wpływ innych zmiennych.
- **Nie spełnione jest założenie o niezależności obserwacji: ignorujemy fakt pochodzenia zębów od tego samego osobnika.**

Wątpliwości i zarzuty

- Czy model jest adekwatny?
- Wpływ innych zmiennych.
- **Nie spełnione jest założenie o niezależności obserwacji: ignorujemy fakt pochodzenia zębów od tego samego osobnika.**

Wątpliwości i zarzuty

- Czy model jest adekwatny?
- Wpływ innych zmiennych.
- **Nie spełnione jest założenie o niezależności obserwacji: ignorujemy fakt pochodzenia zębów od tego samego osobnika.**

Dane zgrupowane

120 osobników, 5 cech (zmiennych).

Cecha wyjaśniana: obecność zębiniaków zapisana w postaci (liczba zębów z zębinikami, liczba zębów bez zębiniaka) = (zeb 1, zeb 0).

Osobnik	plec	wiek	starcie	prochnica	zeb 1	zeb 0
1	1	37	8.00	0.33	0	3
2	1	20	2.83	0.67	2	10
3	1	23	2.86	1.00	4	3
4	1	33	4.40	0.80	0	5
5	0	18	1.00	0.38	1	7
6	1	23	1.88	0.88	1	7
...	...	...	...	...	...	...
12	0	22	2.30	0.90	4	6
...	...	...	...	...	...	...
120	1	23	3.10	1.00	0	10

Trzy modele jednowymiarowe

1. i – numer zęba ($i = 1, \ldots, 780$).

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1,$$

   - x_i – zmienna „wyjaśniająca”: starcie.
   - Przynależność zębów do osobników jest zignorowana. Założenie o niezależności wierszy – niespełnione.

2. j – numer osobnika ($j = 1, \ldots, 120$).

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \bar{x}_j \beta_1,$$

   - $\bar{x}_j$ – zmienna „wyjaśniająca”: średnie starcie.

3. i – numer zęba, j_i – numer osobnika dla i-tego zęba.

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + u_{j_i},$$

   - x_i – zmienna „wyjaśniająca”: starcie.
   - Przynależność zębów do osobników jest uwzględniana. Pojawia się 120 „parametrów zakłócających”.

Trzy modele jednowymiarowe

1. i – numer zęba ($i = 1, \ldots, 780$).

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1,$$

   - x_i – zmienna „wyjaśniająca”: starcie.
   - Przynależność zębów do osobników jest zignorowana. Założenie o niezależności wierszy – niespełnione.

2. j – numer osobnika ($j = 1, \ldots, 120$).

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \bar{x}_j \beta_1,$$

   - $\bar{x}_j$ – zmienna „wyjaśniająca”: średnie starcie.

3. i – numer zęba, j_i – numer osobnika dla i-tego zęba.

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + u_{j_i},$$

   - x_i – zmienna „wyjaśniająca”: starcie.
   - Przynależność zębów do osobników jest uwzględniana. Pojawia się 120 „parametrów zakłócających”.

Trzy modele jednowymiarowe

1. i – numer zęba ($i = 1, \ldots, 780$).

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1,$$

   - x_i – zmienna „wyjaśniająca”: starcie.
   - Przynależność zębów do osobników jest zignorowana. Założenie o niezależności wierszy – niespełnione.

2. j – numer osobnika ($j = 1, \ldots, 120$).

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \bar{x}_j \beta_1,$$

   - $\bar{x}_j$ – zmienna „wyjaśniająca”: średnie starcie.

3. i – numer zęba, j_i – numer osobnika dla i-tego zęba.

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + u_{j_i},$$

   - x_i – zmienna „wyjaśniająca”: starcie.
   - Przynależność zębów do osobników jest uwzględniana. Pojawia się 120 „parametrów zakłócających”.

Trzy modele jednowymiarowe

1. i – numer zęba ($i = 1, \ldots, 780$).

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1,$$

   - x_i – zmienna „wyjaśniająca”: starcie.
   - Przynależność zębów do osobników jest zignorowana. Założenie o niezależności wierszy – niespełnione.

2. j – numer osobnika ($j = 1, \ldots, 120$).

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \bar{x}_j \beta_1,$$

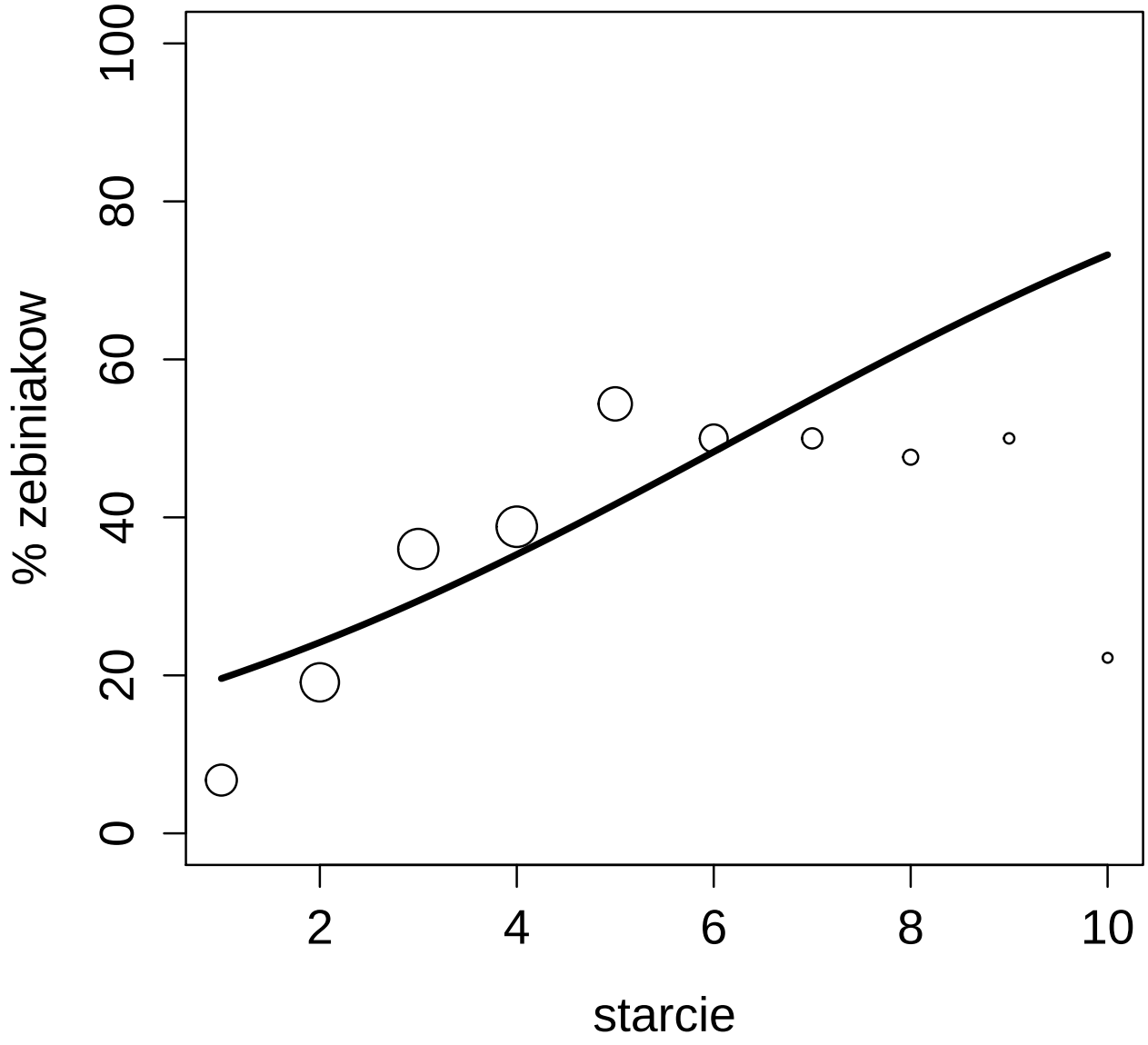
   - $\bar{x}_j$ – zmienna „wyjaśniająca”: średnie starcie.

3. i – numer zęba, j_i – numer osobnika dla i-tego zęba.

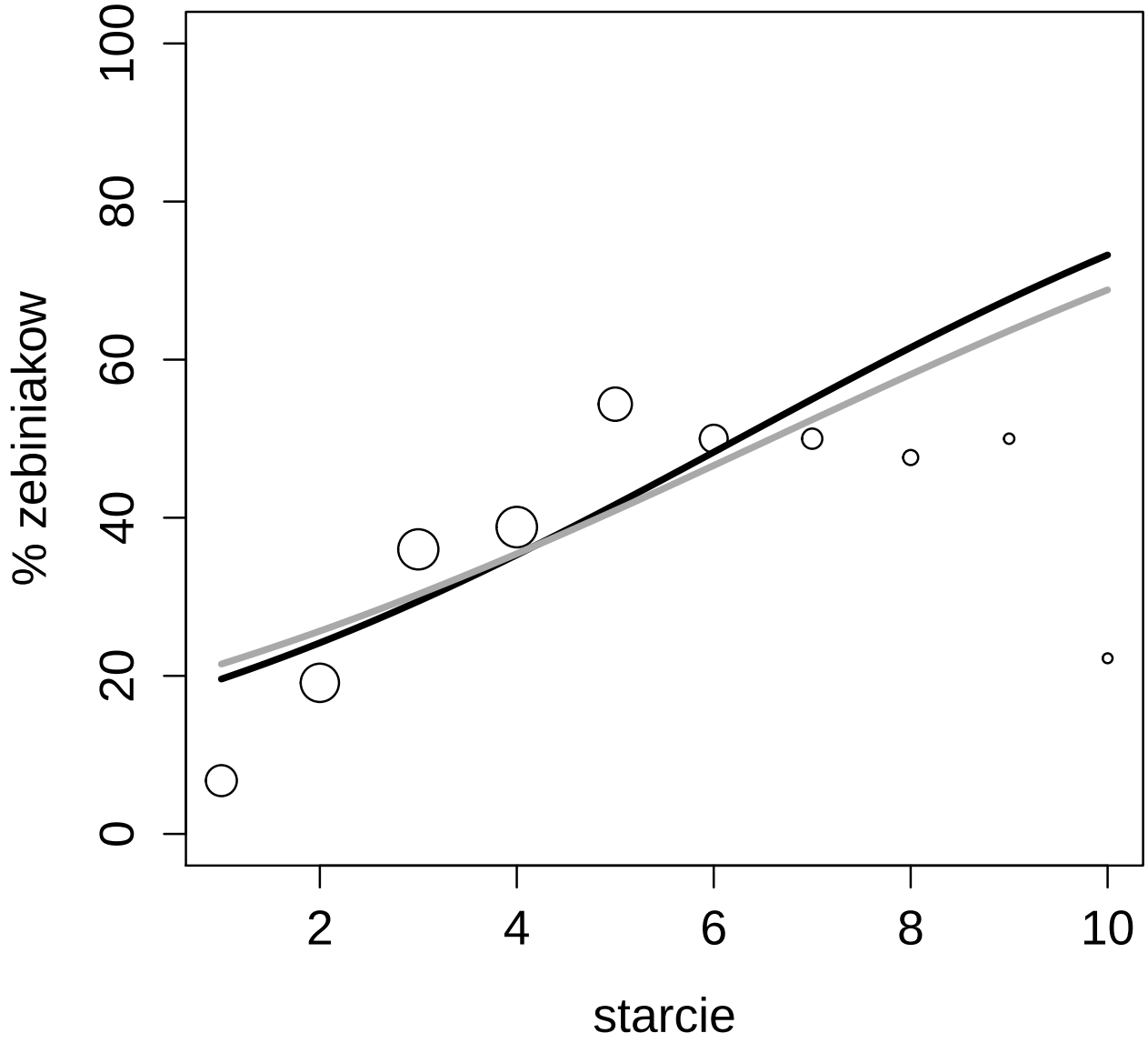
$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + u_{j_i},$$

   - x_i – zmienna „wyjaśniająca”: starcie.
   - Przynależność zębów do osobników jest uwzględniana. Pojawia się 120 „parametrów zakłócających”.

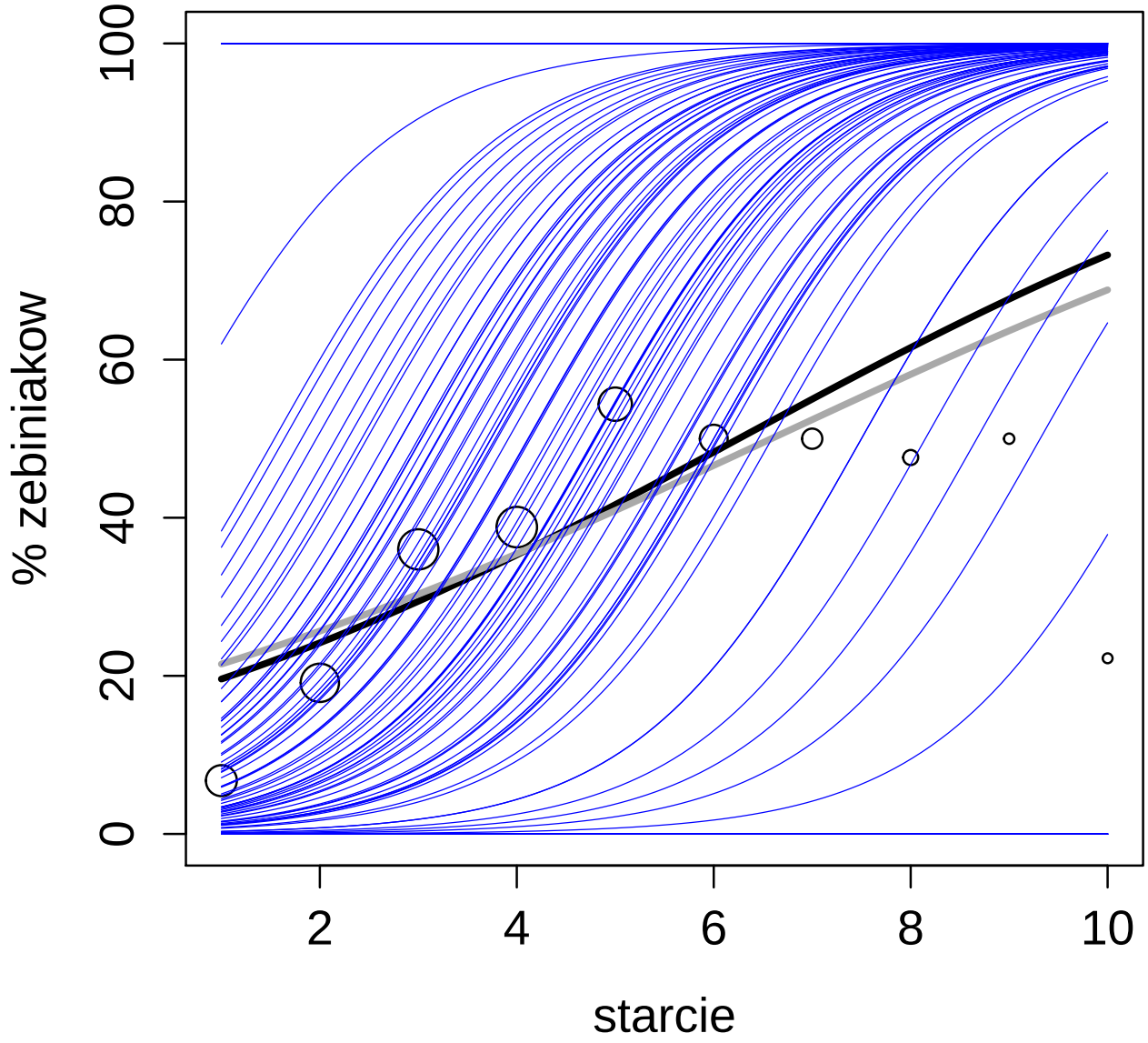
Regresja logistyczna: model 1



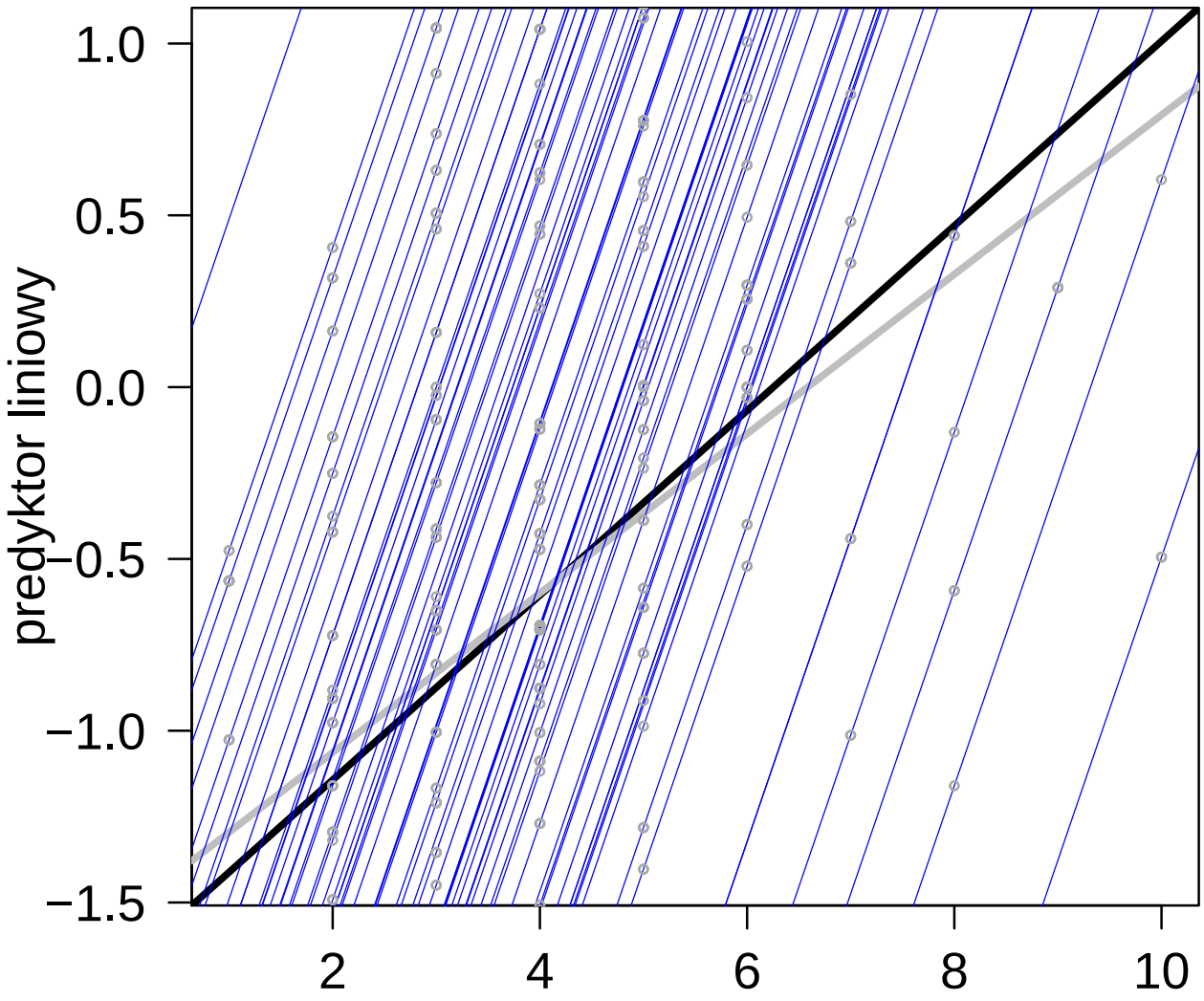
Regresja logistyczna: modele 1 i 2



Regresja logistyczna: modele 1 i 2 i 3



Predyktory liniowe dla trzech modeli jednowymiarowych



Czwarty model jednowymiarowy

- Efekty stałe czy efekty losowe: *Generalized Linear Models* (GLM) vs *Generalized Linear Mixed Models* (GLMM) ?

  i – numer zęba. j_i – numer osobnika dla i-tego zęba.

$$\log \frac{p_i}{1-p_i} = \beta_0 + x_i\beta_1 + u_{j_i},$$

3. Efekty stałe (GLM):

   „Efekt osobnika” u_j jest traktowany jako nieznany parametr. Pojawia się 120 „parametrów zakłócających” u_j.

4. Efekty losowe (GLMM):

   „Efekt osobnika” u_j jest traktowany jako zmienna losowa, $u_j \sim N(0, \sigma_u^2)$, niezależne dla $j = 1, \ldots, 120$.

Czwarty model jednowymiarowy

- Efekty stałe czy efekty losowe: *Generalized Linear Models* (GLM) vs *Generalized Linear Mixed Models* (GLMM) ?

  i – numer zęba. j_i – numer osobnika dla i-tego zęba.

$$\log \frac{p_i}{1-p_i} = \beta_0 + x_i\beta_1 + u_{j_i},$$

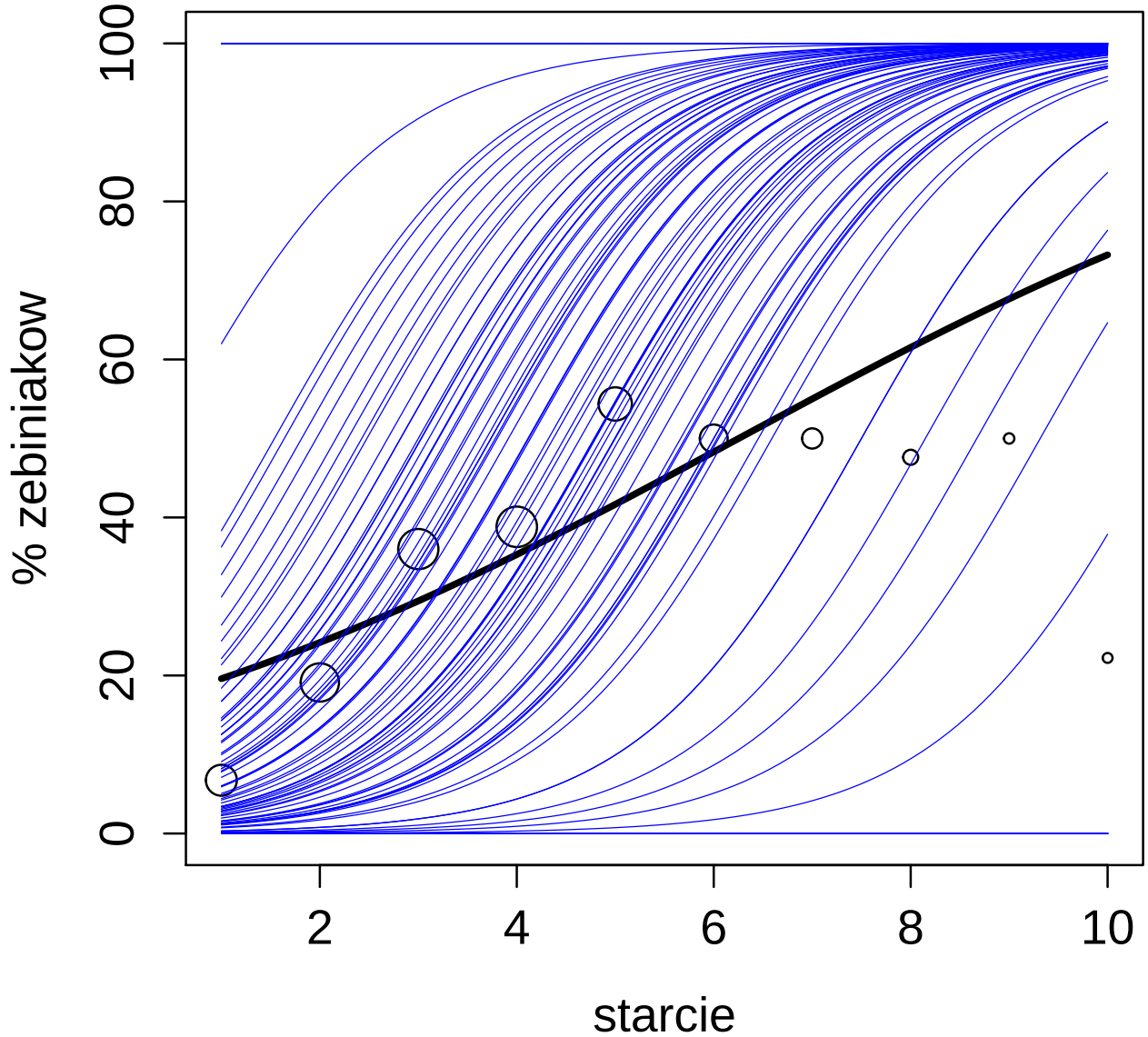
3 Efekty stałe (GLM):

„Efekt osobnika” u_j jest traktowany jako nieznany parametr. Pojawia się 120 „parametrów zakłócających” u_j.

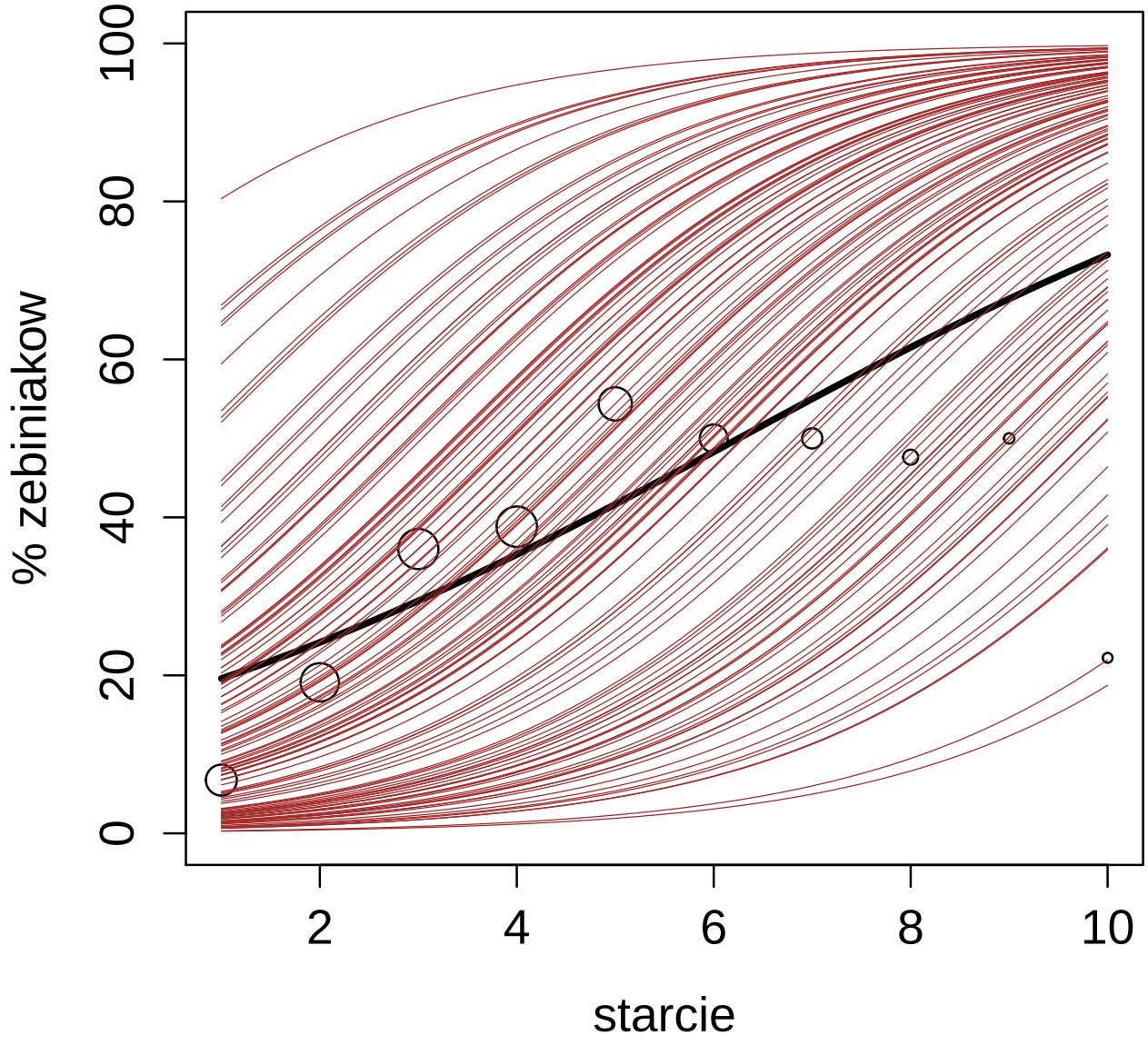
4 Efekty losowe (GLMM):

„Efekt osobnika” u_j jest traktowany jako zmienna losowa, $u_j \sim N(0, \sigma_u^2)$, niezależne dla $j = 1, \ldots, 120$.

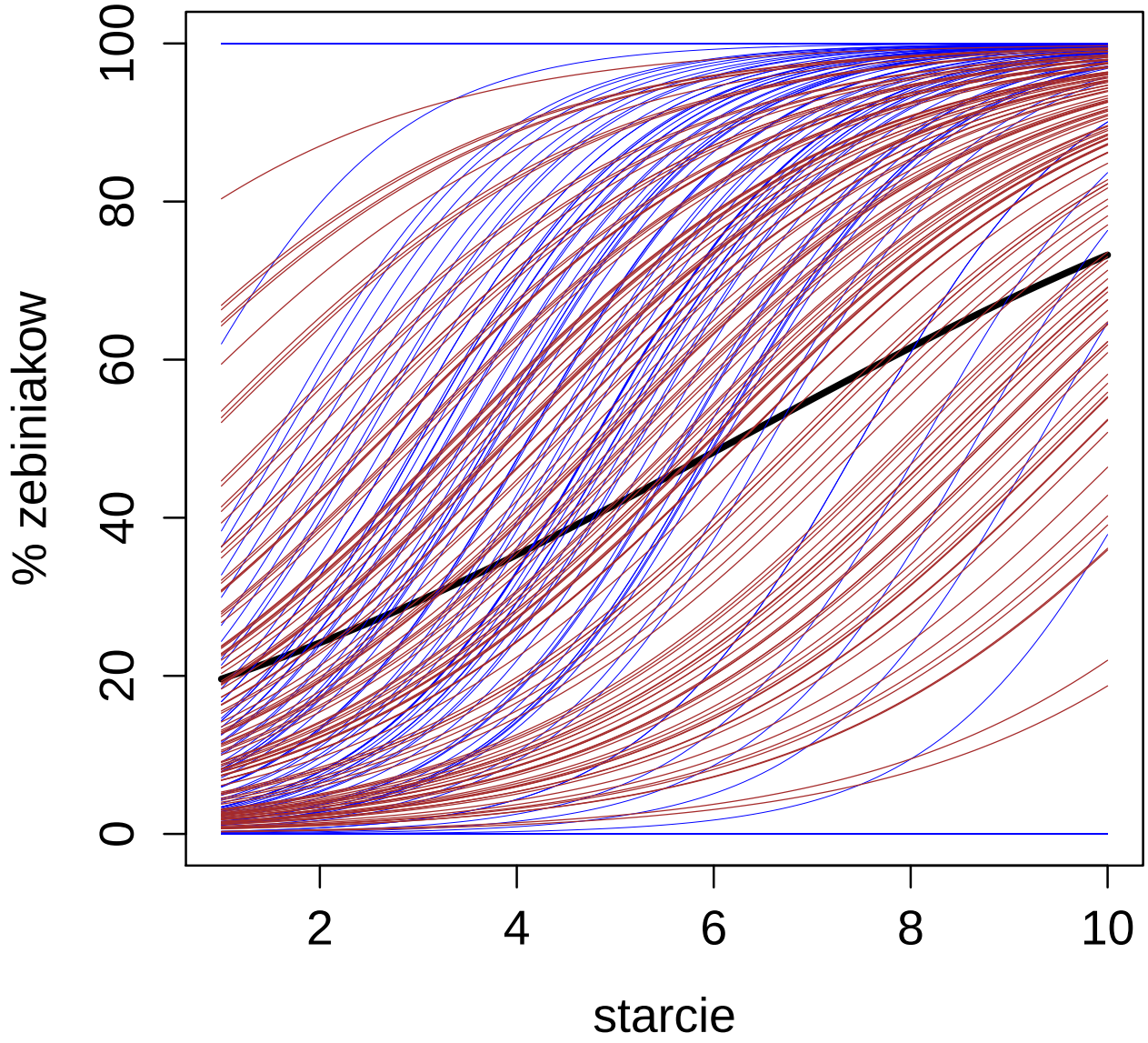
Regresja logistyczna: model 3



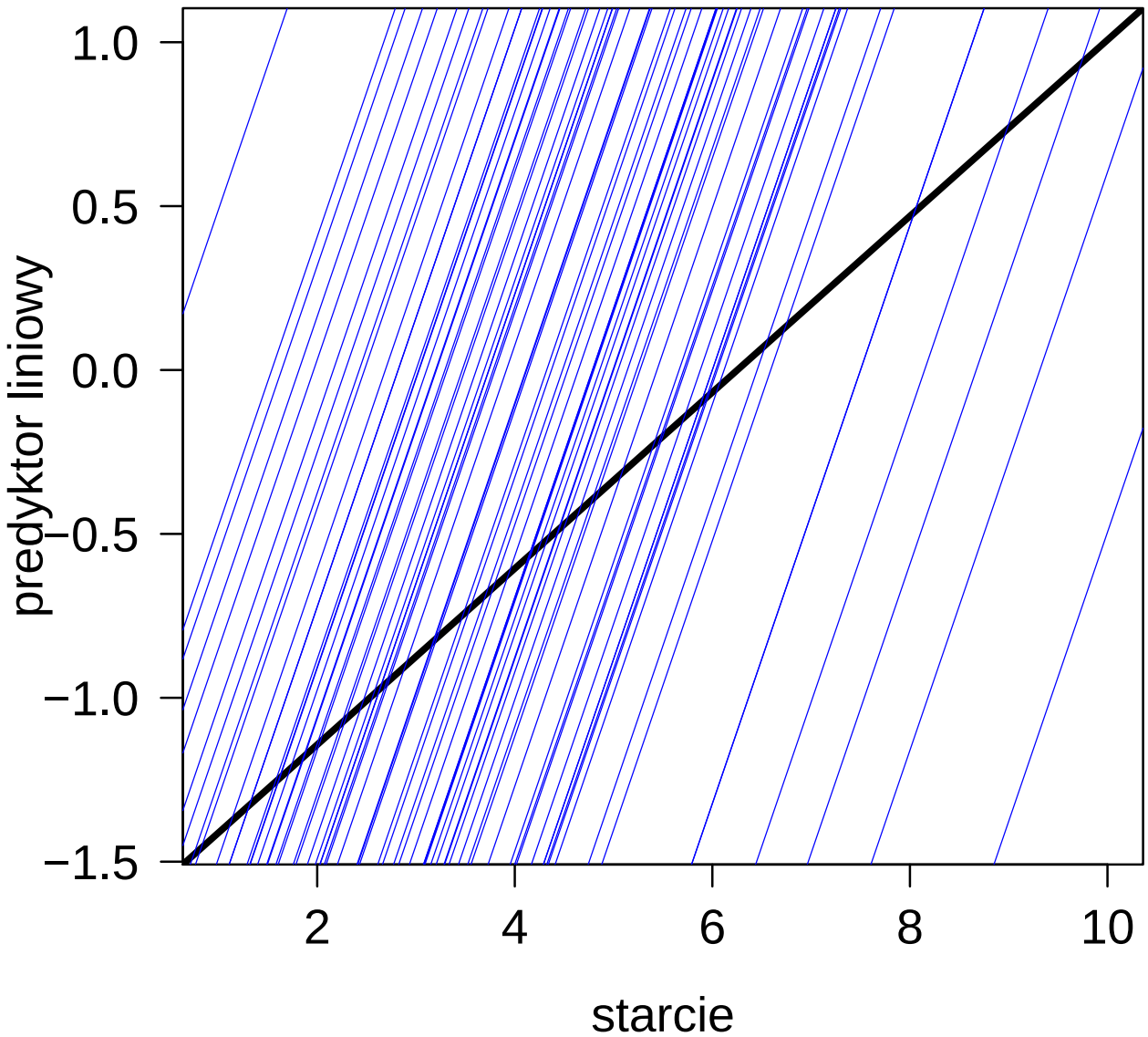
Regresja logistyczna: model 4



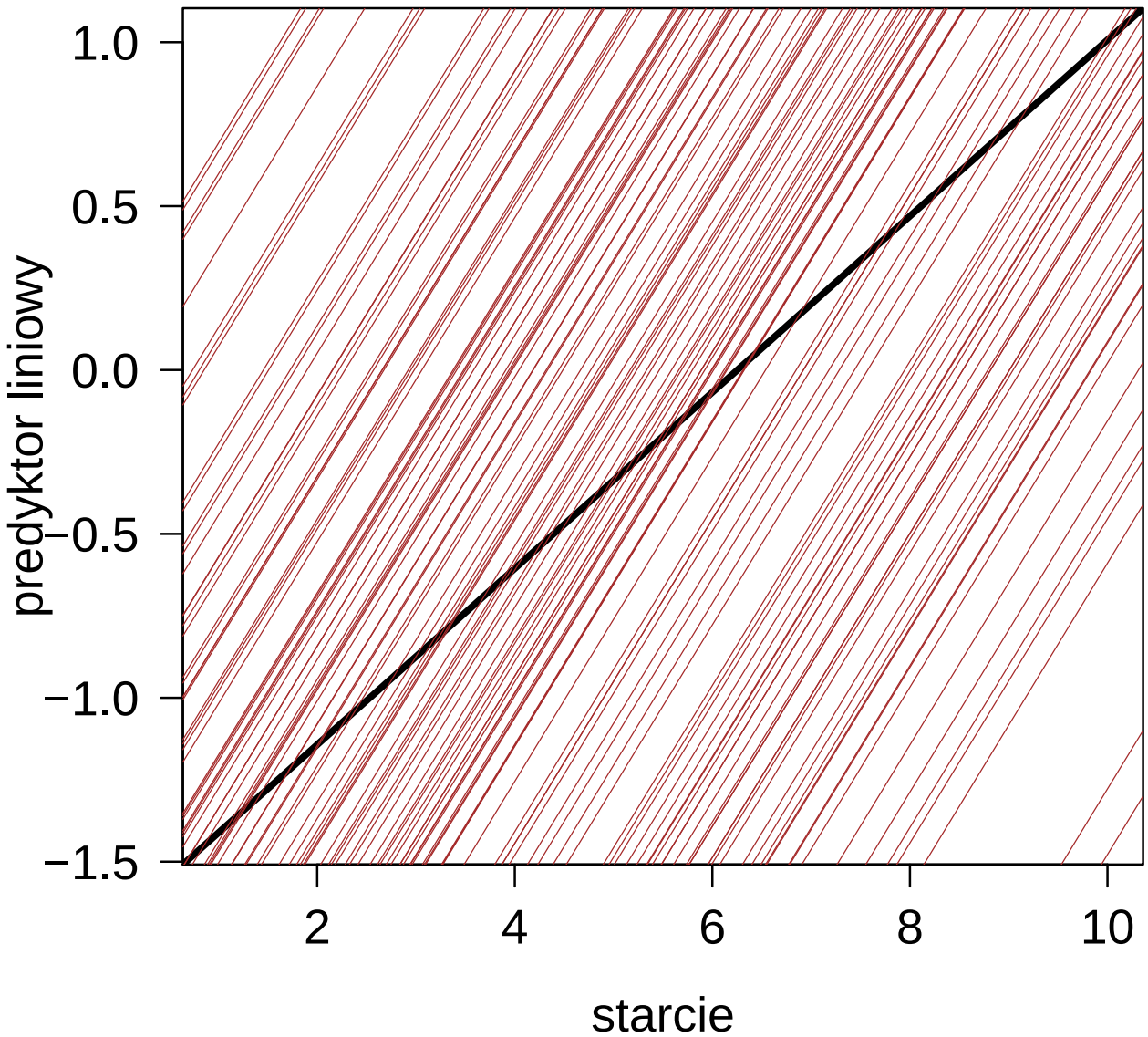
Regresja logistyczna: modele 3 i 4



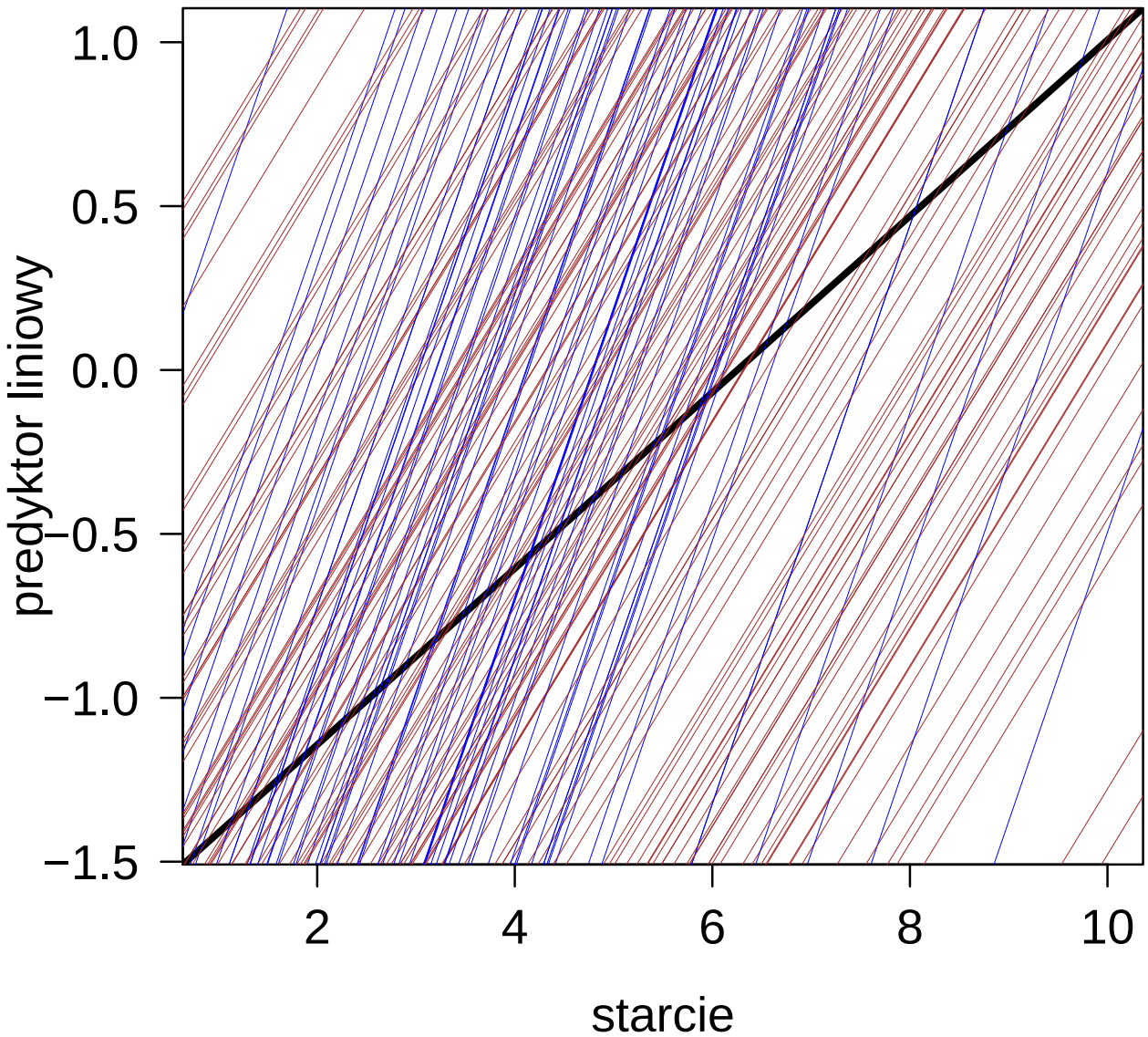
Predyktory liniowe: model 3



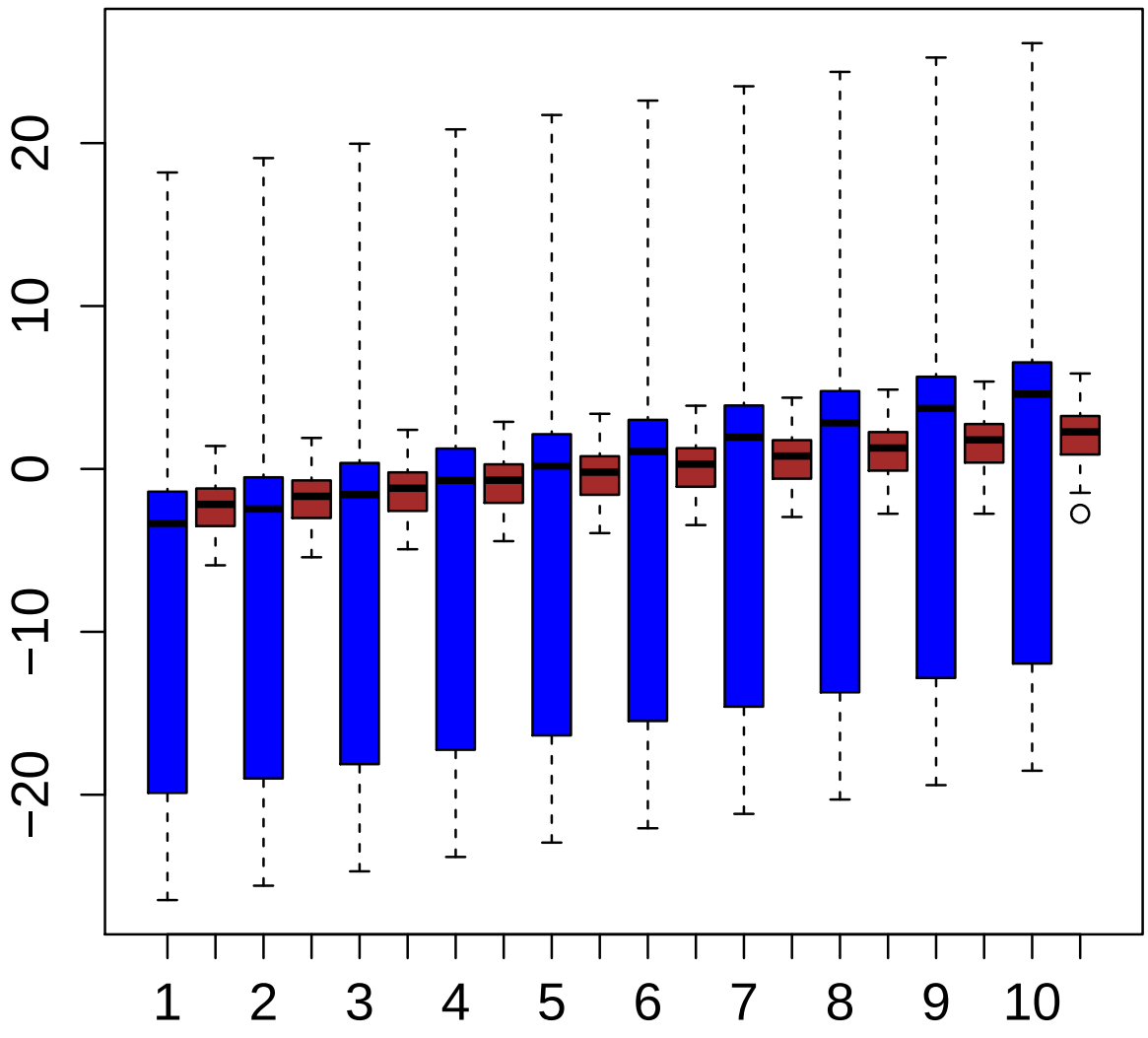
Predyktory liniowe: model 4



Predyktory liniowe: modele 3 i 4



Predyktory liniowe: modele 3 i 4



Cztery modele jednowymiarowe kwadratowo-logistyczne

1^2 i – numer zęba ($i = 1, \ldots, 780$).

$$\log \frac{p_i}{1-p_i} = \beta_0 + x_i\beta_1 + x_i^2\beta_2,$$

2^2 j – numer osobnika ($j = 1, \ldots, 120$).

$$\log \frac{p_j}{1-p_j} = \beta_0 + \bar{x}_j\beta_1 + \bar{x}_j^2\beta_2,$$

- Model zastosowany do zgrupowanych danych.

3^2 i 4^2 i – numer zęba, j_i – numer osobnika dla i-tego zęba.

$$\log \frac{p_i}{1-p_i} = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + u_{j_i},$$

3^2 Efekt osobnika u_j - nielosowy parametr.

4^2 Efekt osobnika - losowy $u_j \sim N(0, \sigma_u^2)$.

Cztery modele jednowymiarowe kwadratowo-logistyczne

1^2 i – numer zęba ($i = 1, \ldots, 780$).

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + x_i^2 \beta_2,$$

2^2 j – numer osobnika ($j = 1, \ldots, 120$).

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \bar{x}_j \beta_1 + \bar{x}_j^2 \beta_2,$$

- Model zastosowany do zgrupowanych danych.

3^2 i 4^2 i – numer zęba, j_i – numer osobnika dla i-tego zęba.

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + x_i^2 \beta_2 + u_{j_i},$$

3^2 Efekt osobnika u_j - nielosowy parametr.

4^2 Efekt osobnika - losowy $u_j \sim N(0, \sigma_u^2)$.

Cztery modele jednowymiarowe kwadratowo-logistyczne

1^2 i – numer zęba ($i = 1, \ldots, 780$).

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + x_i^2 \beta_2,$$

2^2 j – numer osobnika ($j = 1, \ldots, 120$).

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \bar{x}_j \beta_1 + \bar{x}_j^2 \beta_2,$$

- Model zastosowany do zgrupowanych danych.

3^2 i 4^2 i – numer zęba, j_i – numer osobnika dla i-tego zęba.

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + x_i^2 \beta_2 + u_{j_i},$$

3^2 Efekt osobnika u_j - nielosowy parametr.

4^2 Efekt osobnika - losowy $u_j \sim N(0, \sigma_u^2)$.

Cztery modele jednowymiarowe kwadratowo-logistyczne

1^2 i – numer zęba ($i = 1, \ldots, 780$).

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + x_i^2 \beta_2,$$

2^2 j – numer osobnika ($j = 1, \ldots, 120$).

$$\log \frac{p_j}{1 - p_j} = \beta_0 + \bar{x}_j \beta_1 + \bar{x}_j^2 \beta_2,$$

- Model zastosowany do zgrupowanych danych.

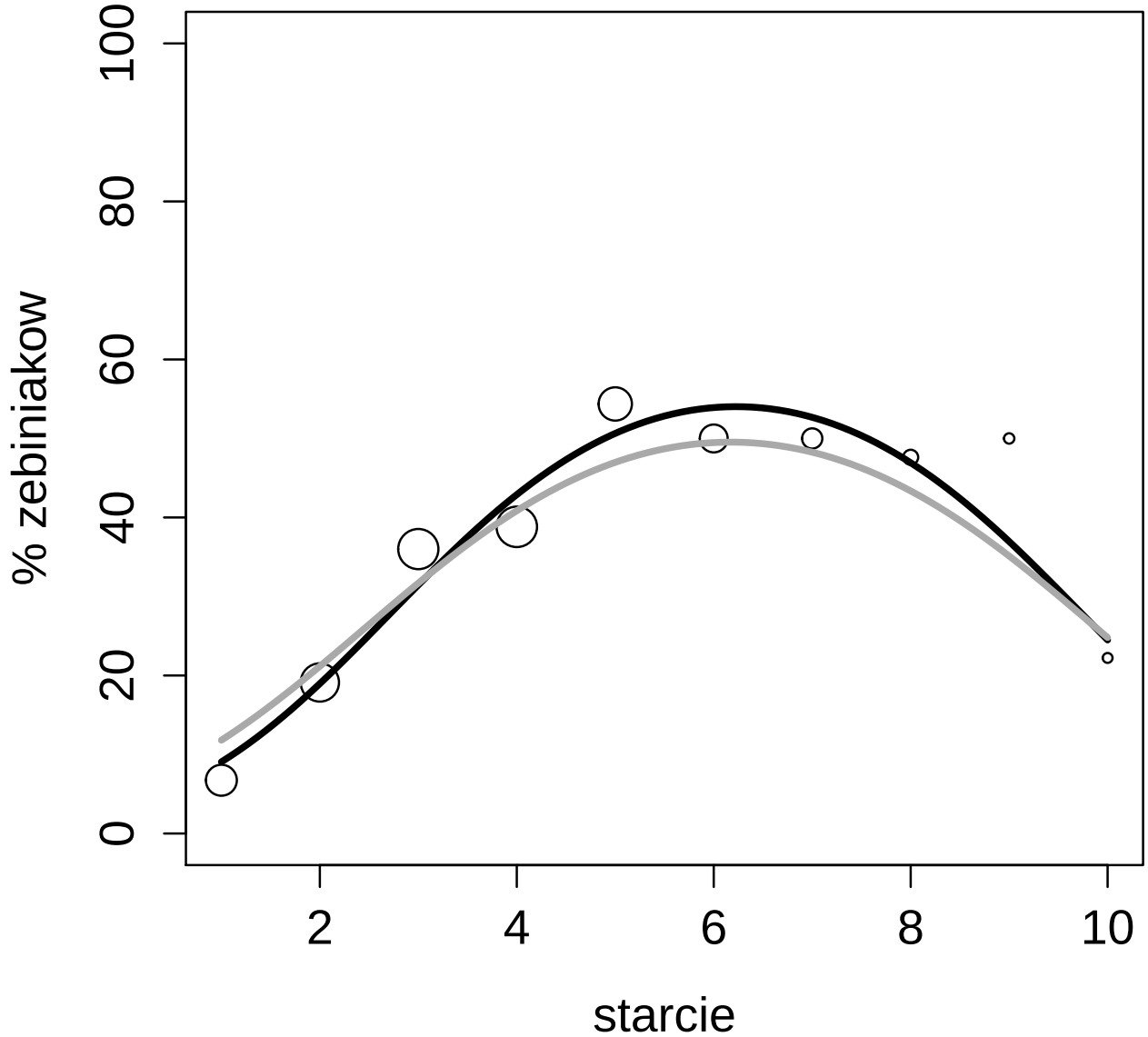
3^2 i 4^2 i – numer zęba, j_i – numer osobnika dla i-tego zęba.

$$\log \frac{p_i}{1 - p_i} = \beta_0 + x_i \beta_1 + x_i^2 \beta_2 + u_{j_i},$$

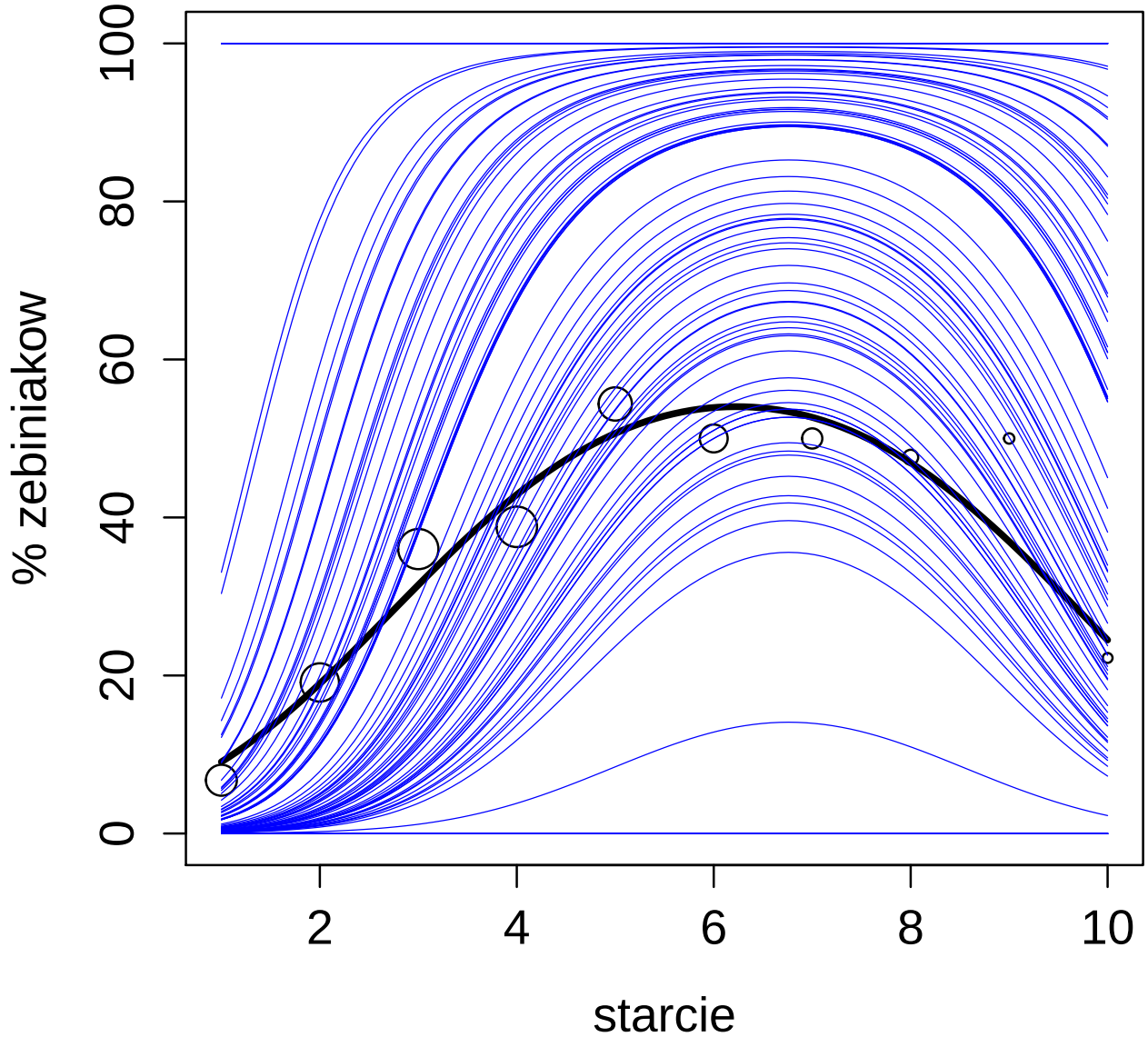
3^2 Efekt osobnika u_j - nielosowy parametr.

4^2 Efekt osobnika - losowy $u_j \sim N(0, \sigma_u^2)$.

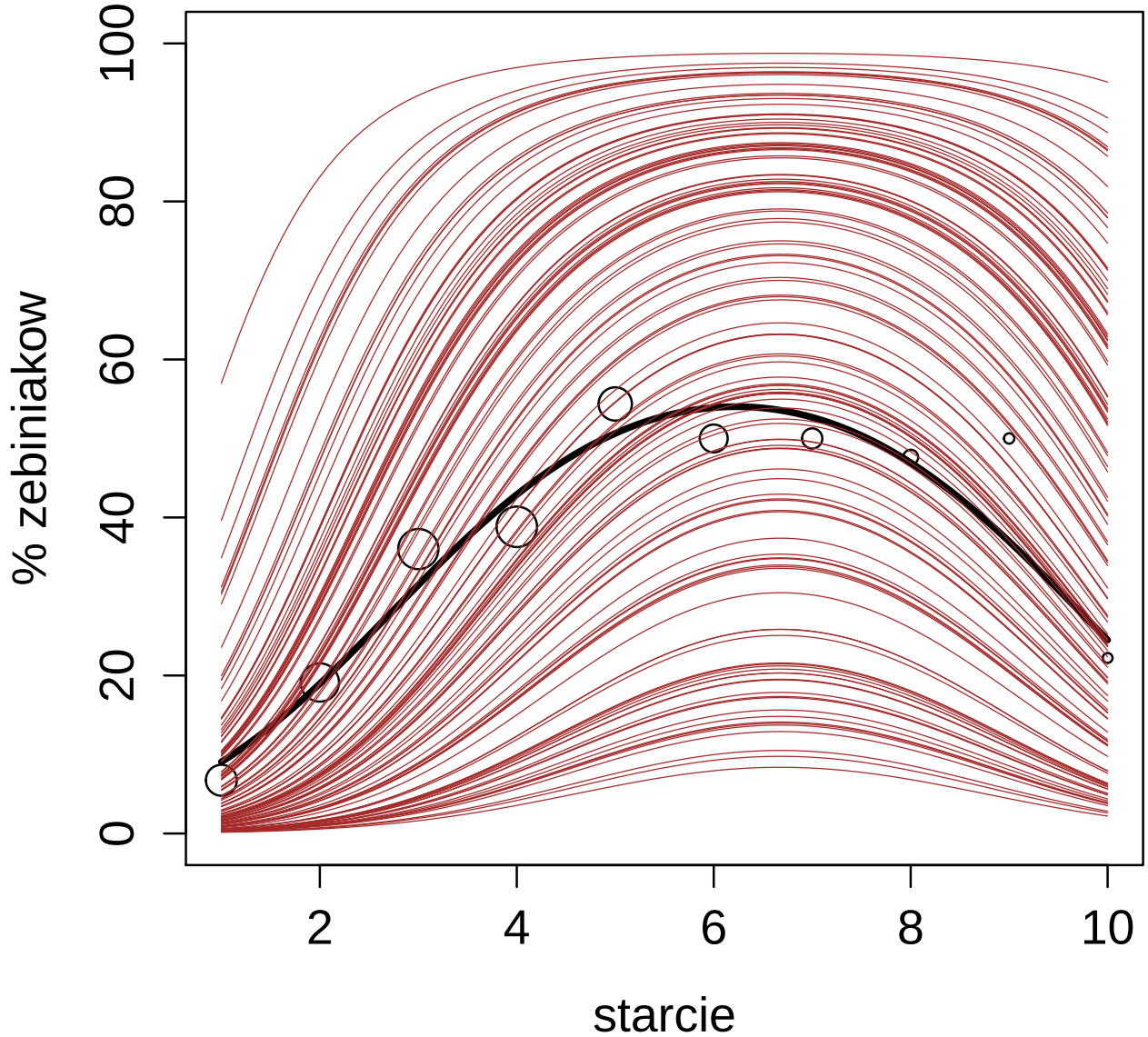
Regresja logistyczna: modele 1^2 i 2^2



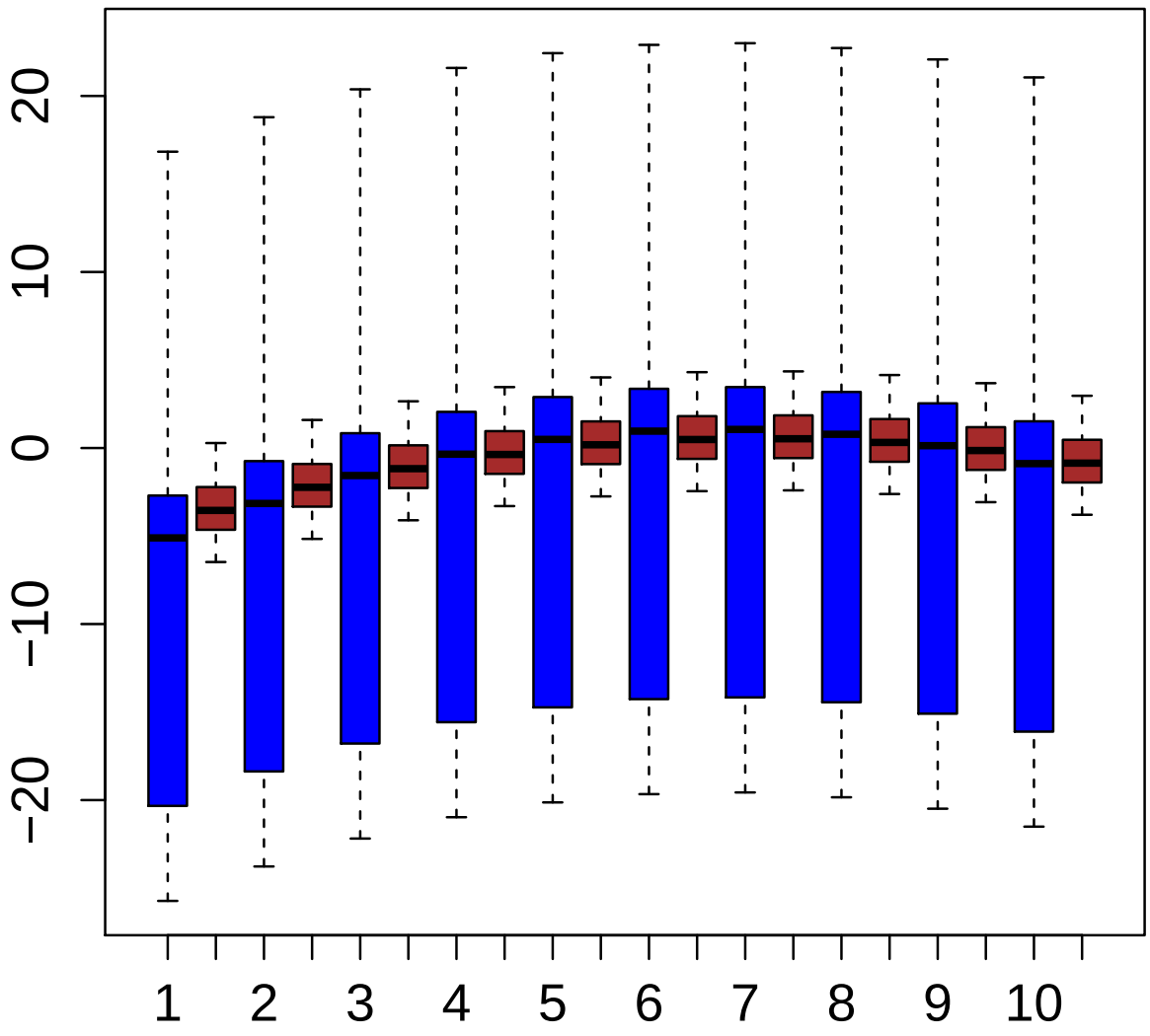
Regresja logistyczna: modele 1^2 i 3^2



Regresja logistyczna: modele 1^2 i 4^2



Predyktory kwadratowe: modele 3^2 i 4^2



Regresja wieloraka (wiele zmiennych objaśniających)

Model 1^4: Dane indywidualne: 780 zębów. Przynależność do grup (osobników) - ignorowana.

Zmienne objaśniające: starcie, starcie^2, wiek, prochnica.

Cecha wyjaśniana: obecność zębiniaków.

Call:
glm(formula = zebiniak $\sim$ starcie + starcie2 + wiek + prochnica, family = binomial, data = c)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.80158	0.47564	-7.993	1.32e-15	***
starcie	1.06419	0.17767	5.990	2.10e-09	***
starcie2	-0.08585	0.01697	-5.060	4.20e-07	***
wiek	0.00749	0.01377	0.544	0.5865	
prochnica	0.45931	0.23424	1.961	0.0499	*

Regresja wieloraka (wiele zmiennych objaśniających)

Model 2^4: Dane zgrupowane: 120 osobników.

Zmienne objaśniające: starcie, starcie^2, wiek, prochnica.

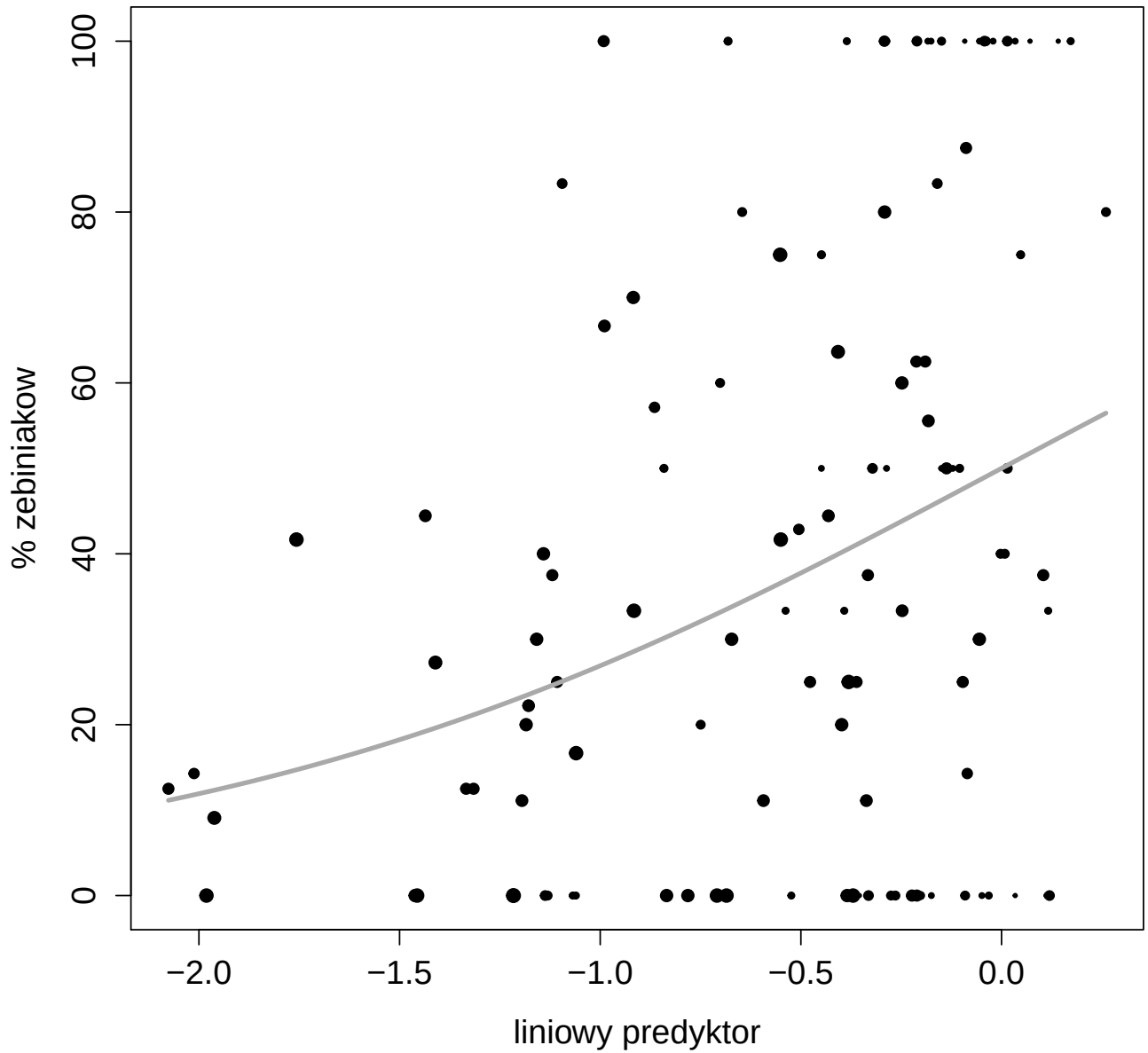
Cecha wyjaśniana: obecność zębiniaków.

Call:
glm(formula = cbind(zeb1, zeb0) $\sim$ starcie + starcie2 + wiek + prochnica, family = binomial, data = c)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-3.32058	0.51797	-6.411	1.45e-10	***
starcie	0.71226	0.22148	3.216	0.00130	**
starcie2	-0.06221	0.02104	-2.957	0.00311	**
wiek	0.02627	0.01594	1.648	0.09945	.
prochnica	0.32465	0.39873	0.814	0.41552	

Regresja wieloraka: model 2^4



Regresja wieloraka (wiele zmiennych objaśniających)

Model 4^4: Dane indywidualne: 780 zębów. Przynależność do grup (osobników) - modelowana jako **efekty losowe**.

Zmienne objaśniające: starcie, starcie^2, wiek, prochnica. Efekty losowe: Osobnik.

Cecha wyjaśniana: obecność zębiniaków.

Linear mixed-effects model fit by maximum likelihood
Fixed effects: zebiniak $\sim$ starcie + starcie2 + wiek + prochnica
Number of Observations: 780
Number of Groups: 120

Coefficients:

	Estimate	Std. Error	t-value	p-value	
(Intercept)	-5.413214	0.8908634	-6.076368	0.0000	***
starcie	1.721269	0.2491799	6.907737	0.0000	***
starcie2	-0.126093	0.0247294	-5.098887	0.0000	***
wiek	-0.014568	0.0259776	-0.560791	0.5760	
prochnica	0.815126	0.2786736	2.925019	0.0036	**


Julian J. Faraway, *Extending the Linear Model with R. Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall 2006.


P. McCullagh and J.A. Nelder, *Generalized Linear Models*, SECOND EDITION. Chapman & Hall 1989.


Przemysław Biecek, *Analiza danych z programem R. Modele liniowe z efektami stałymi, losowymi i mieszanymi*. Wydawnictwo Naukowe PWN 2011.


Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.